图书情报工作
第 62 卷 第 22 期　2018 年 11 月

地方政府开放数据的评估框架与发现*

■ 郑磊　吕文增
复旦大学国际关系与公共事务学院　上海 200433

摘要：[**目的/意义**] 构建针对我国地方政府开放数据的评估框架，对现有的地方政府开放数据平台上的数据层面进行评价并提出建议，以助推地方政府数据开放。[**方法/过程**] 根据政府数据开放的定义、原则与标准，借鉴国际开放数据评估框架，基于目前我国政府数据开放的政策要求和发展现状，汇聚各界专家学者的意见，构建起一个系统科学、多维度、可操作的政府数据开放评估框架，并基于该框架对我国现有的 46 个地方政府开放数据平台上的数据进行综合评估。[**结果/结论**] 研究发现我国地方政府开放的数据在数量、质量、标准、覆盖面和可持续性方面存在的各类问题。

关键词：中国　地方政府　数据开放　评估　框架
分类号：G250
DOI：10.13266/j.issn.0252-3116.2018.22.004

1　引言

政府部门在履行行政职责过程中制作、获取和保存的数据资源是整个社会的公共资源，在保障国家秘密、商业秘密和个人隐私的前提下，将政府数据最大限度地开放给社会进行开发利用，将有利于提升政府透明度，激发创新创业活力，转变经济发展方式，提高公共服务水平，提升政府治理能力[1]。近年来，随着开放政府数据在全球范围内的迅速推进，我国政府高度重视开放政府数据。2012 年以来，我国已有近 50 个地方政府陆续推出数据开放平台，取得了一定成效，也积累了不少经验。然而，我国地方政府到底开放了多少数据，这些数据的标准和质量如何，覆盖了那些领域？是否可持续开放？还存在哪些问题和挑战，亟需开展深入评测和研究。

本文首先构建起一个系统科学、多维度、可操作的开放数据评估框架，并基于该框架对我国现有的地方政府开放数据平台上的数据层面进行综合评价，提出优化和提升建议，希望有助于我国地方政府数据开放的推进与发展。

2　文献综述

2.1　关于政府数据开放的基本原则和标准

2007 年 12 月，30 位开放数据倡导者聚集在美国加州举行会议，共同提出了政府数据开放的 8 项基本原则[2]：完整的、一手的、及时的、可获取的、可机读的、对非歧视性的、非专属的、免授权的。根据世界银行的定义，开放数据是指“能被任何人出于任何目的不受限制地进行自由利用、再利用和分发，并最大程度保持其原始出处和开放性的数据”。开放定义指出开放意味着任何人都可以出于任何目的自由地访问、使用、修改和共享数据[3]。“开放性”应具备两个维度的特性：一为技术性开放，即数据应为可机读、非专属性的电子格式，从而能被任何人使用和通用、能被免费的软件获取和利用，数据还应被置于公共服务器上供公众获取，不设密码和防火墙；二为法律性开放，即这些数据必须被置于公共领域，或处于自由利用条款下，受到最低程度的限制[4]。2010 年，万维网的发明人、语义网和关联数据的创建者和倡导者 T. Berners-Lee 提出了一个开放数据五星标准[5]：一星是指基于开放授权在网络上开放数据，用户可以查看、搜索、存储和修改数据，还可以与任何人分享这些数据，但对数据格式不做要求，可能采用 PDF、JPEG 等格式；二星是指以可机读、结构化格式开放数据，例如 EXCEL 电子表格的形式，但不包括表格的图像扫描件；三星是指在满足二星标准的基础上，以非专属开放格式开放数据，如采用 CSV 格式

* 本文系国家自然科学基金项目“大数据背景下开放政府数据的因素与机理研究：系统动力学建模与政策仿真”（项目编号：71473048）研究成果之一。
作者简介：郑磊（ORCID：0000-0002-8549-7428），副教授，博士，硕士生导师，E-mail：zhengl@fudan.edu.cn；吕文增（ORCID：0000-0001-9602-5859），硕士研究生。
收稿日期：2018-08-18　修回日期：2018-10-09　本文起止页码：32-44　本文责任编辑：王传清

而不是 EXCEL 格式，用户不需要使用专属的、付费的软件就可以分析数据；四星是指在满足以上要求的基础上，采用 W3C 开放标准的数据（如 RDF 和 SPARQL 格式），为每一个数据集设置固定的 URL 链接，便于使用者发现和链接到数据集的具体位置；五星是指在满足以上要求的基础上，借助 W3C 标准和关联数据原则，使数据之间实现关联，提供数据的背景。2015 年，《开放数据宪章》将开放数据界定为具备必要的技术和法律特性，从而能被任何人、在任何时间和地点进行自由利用、再利用和分发的电子数据。该宪章还提出了政府数据开放应遵循的六大原则[6]：默认开放、及时和全面、可获取和可利用、可比较和互操作性、致力于改善治理和公民参与、致力于包容性发展和创新。

我国对于政府数据开放的政策要求也与以上国际标准相符。2017 年 2 月，中央全面深化改革领导小组审议通过的《关于推进公共信息资源开放的若干意见》指出，要保证开放数据的"完整性、准确性、原始性、机器可读性、非歧视性、及时性，方便公众在线检索、获取和利用"。2017 年 5 月，国务院办公厅印发的《政务信息系统整合共享实施方案》指出，要向社会开放"政府部门和公共企事业单位的原始性、可机器读取、可供社会化再利用的数据集"。2018 年 1 月，中央网信办、发展改革委以及工业和信息化部联合印发的《公共信息资源开放试点工作方案》要求试点地区"提升数据的完整性、准确性、有效性、时效性"，"明确开放数据的完整性、机器可读性、格式通用性等要求"。国内外相关机构和专家对政府数据开放标准的梳理结果如表 1 所示：

表 1　政府数据开放的标准

机构/文件	标准
政府数据开放 8 项基本原则	完整的、一手的、及时的、可获取的、可机读的、非歧视性的、非专属的、免授权的
世界银行	"技术性开放"和"法律性开放"
T. Berners-Lee 的"开放数据五星标准"	开放授权、可机读、结构化、非专属性、W3C 开放标准、关联数据
《开放数据宪章》	默认开放、及时和全面、可获取和可利用、可比较和互操作性
中央全面深化改革领导小组《关于推进公共信息资源开放的若干意见》	完整性、准确性、原始性、机器可读性、非歧视性、及时性，方便公众在线检索、获取和利用
中央网信办、发展改革委以及工业和信息化部联合印发的《公共信息资源开放试点工作方案》	完整性、准确性、有效性、时效性、机器可读性、格式通用性

2.2　关于政府数据开放评估的研究

E. Oviedo 等[7]建立了一个开放数据平台质量模型，包括可用性、再利用的能力、关联性、可靠性、颗粒度和可视化 6 个维度。G. Viscusi 等[8]分别就完整性、准确性和及时性 3 个数据质量维度，提出了一个基于质量的开放政府数据完成度评估框架。R. P. Lourenço[9]提出了数据质量、平台数据主体和时间的完整性、数据获取的便捷性、数据的可用性和可理解性、及时性、数据价值和有用性、颗粒度 7 个指标。O. Bello 等[10]使用的评估变量包括"五星标准"、实施技术、数据格式、开放许可、关键数据集和功能性。

2014 年，纽约大学治理实验室对国际上具有代表性的十一个研究机构、评估指标、政府部门和咨询公司界定的"开放数据"定义进行梳理后发现，被提及最多的开放数据标准包括免费、公开提供、非排他性、可利用结构、开放授权和可再利用等要求[11]。I. Susha 等[12]围绕元数据、元方法、元理论 3 个维度，对 5 个开放政府数据评估项目进行了比较研究。

我国学者也对政府数据开放的评估方法开展了研究。夏义堃对 7 个国际组织开放政府数据评估项目的评估主题、评估侧重点、评估对象和评估方法进行了比较和总结[13]。郑磊和关文雯通过对 11 个具代表性的国内外评估项目的评估框架、指标和方法进行梳理分析后发现，目前开放政府数据评估项目的指标体系主要包含基础、平台、数据、使用和效果 5 个维度，而重点是数据和基础两个层面[14]。郑跃平和刘美岑对世界银行的"开放数据准备度"、万维网基金的"开放数据晴雨表"、开放知识基金的"全球开放数据指数"、经济合作与发展组织的"OURdata 指数"以及联合国的"开放政府数据调查"等国外几个具有代表性的开放数据评估项目，从起始时间、评估频率、评估对象、评估工具、数据获取及分值计算等多个维度进行了对比，归纳出这些评估项目的共同点和差异之处，探讨了已有研究存在的一些问题和不足[15]。陈美利用文献调研和案例分析的研究方法分析了美国、英国、澳大利亚主要国家以及国际组织在开放政府数据价值评估的具体实践，提出了认识开放政府数据价值评估的困境、重视并开展开放政府数据的价值评估、注重开放政府数据价值评估方法等建议[16]。韦忻伶、安小米等对现有的开放政府数据评估体系进行系统梳理，归纳了现有评估体系的评估动因、评估内容、评估方法和相应特点及适用性，发现现有评估体系在城市层面、特定行业和开放数据成熟度评估方面存在局限，并构建了开放政府

图书情报工作
第 62 卷 第 22 期 2018 年 11 月

数据评估动因、评估内容和评估方法的循环迭代检验机制[17]。

还有一些学者对我们的政府数据开放的现状进行了实际评估。郑磊和高丰首次通过基础层、数据层、平台层 3 个层面的 13 个维度对我国 8 个地方政府的开放数据实践进行了评估[18]。郑磊和熊久阳又进一步对我国 13 个地方政府数据开放平台上数据的技术和法律特性进行了研究，覆盖开放数据的数量、种类、格式、获取方式、及时性、开放授权、元数据、浏览量和下载量等维度[19]。夏义堃梳理了国际上具有代表性的政府数据开放评估体系内容，系统分析了不同评估体系对中国政府数据开放情况的基本认知。发现我国在政府数据开放水平、信息法律制度、组织管理体系以及技术架构等方面还存在一定的差距与不足[20]。赵继娣和张罕仑以上海市政府数据开放为例，结合内容分析数据和访谈资料，从开放数据的提供与管理、公民的参与和数据利用情况 3 个维度入手，对地方政府数据开放成效进行了评价，剖析了地方政府数据开放的现状[21]。沈晶等基于政府数据开放平台跨时间纵向发展视角，从开放程度提升度、更新频率兑现提升度、用户利用提升度 3 个维度建立政府数据开放发展速度评估体系，并选取 5 个省级政府数据开放平台爬取数据，得到政府数据开放发展速度指数，同时结合中国开放数林指数，分析了政府数据开放发展态势[22]。海伦和邓崧对我国 13 个城市政府数据开放平台的总体效率、纯技术效率和规模效率进行评估。结果显示，在评估的 13 个城市政府的数据开放网站中有 9 个网站纯技术运行效率相对有效[23]。

2.3 研究现状评述

从以上综述可见，国内外学者对于开放数据评估方法的研究已有一定积累，针对我国政府开放数据开展的评估也已起步。然而，目前国际上关于开放政府数据的评估主要集中于国家层面，而我国学者已开展的针对我们地方政府数据开放平台的评估在样本覆盖上还不不够全面，在评估框架上也不够聚焦。虽然对数据、平台、政策、管理等层面都有所涉及，但专门针对数据这一核心层面开展的评估指标还不够深入和系统。随着我国地方政府数据开放实践的不断推进，越来越多的地方政府推出了数据开放平台，地方政府数据开放的内容和形式更趋多样，亟需针对更大范围的样本，更系统和更聚焦地专门针对数据层开展研究，以全面呈现和深入分析我国地方政府开放数据的现状与问题。

3 评估方法

总体上，本研究依照政府数据开放的原则与标准，参考了国际开放数据评估框架和指标体系，又结合了目前我国政府对于数据开放的政策要求和各地发展现状，吸纳了各界专家学者提出的建议意见，最终确定了研究的评估对象、指标体系和数据采集和分析方法。

3.1 评估框架

本研究重点针对各地数据开放平台上的数据层面进行评估。研究的观察对象为地方政府数据开放平台上可通过直接下载或 API 接口两种方式公开获取的、电子形式的原始数据集及相关信息，不包括未通过公开平台开放，而是通过内部授权、协议开放等形式向社会提供的政府数据。

本研究邀请了国内近 40 位数据领域的专家与学者共同参与构建评估框架。这些专家和学者具有公共管理、信息科学、计算机科学、政治学等不同的学科背景，来自于高校、科研机构、政府和企业，可以反映跨界、多学科、第三方、中立的专业视角和实际需求。首先，专家学者们根据系统、科学、可操作的原则，通过分组讨论提出需要评估的各项指标；然后，再通过全场讨论对各组提出的指标进行梳理归类，合并同类项；之后，全场通过现场投票选出相对重要的指标。由此，初步构建起一个针对中国地方政府数据开放数据层面的评估指标体系，包括数据质量、数据标准、数据可持续性、数据数量和数据覆盖面等一级指标和相应的二级、三级指标如表 2 所示：

表 2 政府数据开放数据层评估框架

一级指标	二级指标	三级指标
数据数量	数据集总量	/
	数据容量	/
数据质量	优质数据	/
	无低质数据	无低容量数据
		无碎片化数据
	无问题数据	无重复创建
		无生硬格式转化
		无无效数据
数据标准	开放授权	/
	技术性开放	可机读格式
		开放格式
		RDF 格式
		API 接口
	元数据完整性	基本元数据覆盖率
		API 描述规范
数据覆盖面	主题覆盖	/
	部门覆盖	/
	高需求关键词覆盖	/
数据可持续性	持续增长	/
	动态更新	/
	历史存档	/

郑磊，吕文增. 地方政府开放数据的评估框架与发现[J]. 图书情报工作,2018,62(22):32－44.

之后，专家学者们再通过在线调查工具匿名对指标的相对重要性进行排序，即将其认为最重要的指标排序为1，其次为2，依次类推。排序结果如表3所示，被排在最重要位置的指标是“数据质量”，这反映了各界对高质量数据集的需求。被排在第2－第5位的指标依次是数据标准、数据可持续性、数据数量和数据覆盖面。

表3　专家学者评估指标相对重要性排序结果

指标名称	排序得分	相对重要性
数据质量	1.40	1
数据标准	2.80	2
数据可持续性	2.97	3
数据数量	3.73	4
数据覆盖面	3.87	5

3.2　评估对象

本研究根据公开报道，以及使用“数据＋开放”“数据＋公开”“公共＋数据”“政务＋数据”“政府＋数据”“地名＋数据”“地名＋政府数据”“地名＋开放数据”等关键词进行搜索，发现截至2018年4月中旬我国已上线的政府数据开放平台，并将符合以下条件的地方政府数据开放平台纳入评估范围：

(1)平台域名中出现gov.cn，作为确定其为政府官方认可的数据开放平台的依据。

(2)平台形式为“统一专有式”或“统一嵌入式”。“统一专有式”是指开放数据统一汇聚在一个专门的平台上进行开放；“统一嵌入式”是指开放数据统一汇聚为一个栏目版块，嵌入在政府门户网站或政务服务网站上。

(3)平台所代表的地方政府的行政级别为地级市以上。

(4)平台上确实开放了电子格式的、可通过下载或接口形式获取的、结构化的数据集。有些名为“数据开放”的平台实质上只提供了非结构化的文本内容或跳转到其他相关网页的链接，不存在可通过下载或接口形式获取的、结构化的数据集。这类平台更多属于传统的“信息公开”门户，因而未被纳入本次评估范围，如新疆维吾尔自治区政务数据开放网、四川省人民政府网站上的“开放数据”模块和广东清远市人民政府网的“数据开放”频道等。

基于以上选择标准，被纳入本研究评估的地方政府数据开放平台共46个，这些平台符合政府数据开放的基本特征，是我国政府数据开放的先行者。具体平台名称、所属地方政府和平台域名如表4所示：

表4　评估范围(按行政层级及拼音首字母排序)

序号	平台名称	地点	层级	平台域名
1	北京市政务数据资源网	北京市	省级	http://www.bjdata.gov.cn
2	开放广东	广东省	省级	http://www.gddata.gov.cn
3	贵州省政府数据开放平台	贵州省	省级	http://www.gzdata.gov.cn
4	江西省政府数据开放网站	江西省	省级	http://data.jiangxi.gov.cn
5	开放宁夏	宁夏回族自治区	省级	http://ningxiadata.gov.cn
6	山东公共数据开放网	山东省	省级	http://data.sd.gov.cn
7	上海政府数据服务网	上海市	省级	http://www.datashanghai.gov.cn
8	浙江政务服务网	浙江省	省级	http://data.zjzwfw.gov.cn
9	广州市政府数据统一开放平台	广东省广州市	副省级	http://www.datagz.gov.cn
10	深圳市政府数据开放平台	广东省深圳市	副省级	http://opendata.sz.gov.cn
11	哈尔滨市数据开放	黑龙江省哈尔滨市	副省级	http://data.harbin.gov.cn
12	武汉政府公开数据服务网	湖北省武汉市	副省级	http://www.wuhandata.gov.cn
13	济南市公共数据开放网	山东省济南市	副省级	http://www.jndata.gov.cn
14	青岛市政府数据开放网	山东省青岛市	副省级	http://data.qingdao.gov.cn
15	宁波市政府数据服务网	浙江省宁波市	副省级	http://www.datanb.gov.cn
16	数据东莞	广东省东莞市	市级	http://dataopen.dg.gov.cn
17	佛山政府数据开放平台	广东省佛山市	市级	http://www.fsdata.gov.cn
18	开放惠州	广东省惠州市	市级	http://data.huizhou.gov.cn
19	开放江门	广东省江门市	市级	http://opendata.jiangmen.gov.cn
20	梅州市人民政府数据开放平台	广东省梅州市	市级	https://www.meizhou.gov.cn/opendata
21	中国阳江数据开放	广东省阳江市	市级	http://www.yangjiang.gov.cn/sjkf

（续表 4）

序号	平台名称	地点	层级	平台域名
22	湛江数据服务网	广东省湛江市	市级	http://data.zhanjiang.gov.cn
23	肇庆数据开放	广东省肇庆市	市级	http://www.zhaoqing.gov.cn/sjkf
24	开放中山	广东省中山市	市级	http://zsdata.zs.gov.cn/web/index
25	贵阳市政府数据开放平台	贵州省贵阳市	市级	http://www.gyopendata.gov.cn
26	荆门市人民政府数据开放模块	湖北省荆门市	市级	http://data.jingmen.gov.cn/app
27	长沙数据开放	湖南省长沙市	市级	http://data.changsha.gov.cn
28	苏州市政府数据开放平台	江苏省苏州市	市级	http://www.suzhou.gov.cn/dataOpenWeb
29	无锡市政府数据服务网	江苏省无锡市	市级	http://etc.wuxi.gov.cn/opendata
30	扬州市政务数据服务网	江苏省扬州市	市级	http://data.yangzhou.gov.cn
31	乌海市数据开放平台	内蒙古自治区乌海市	市级	http://whdata.wuhai.gov.cn/odweb
32	滨州市公共数据开放网	山东省滨州市	市级	http://bzdata.sd.gov.cn
33	德州市公共数据开放网	山东省德州市	市级	http://dzdata.sd.gov.cn
34	东营市公共数据开放网	山东省东营市	市级	http://dydata.sd.gov.cn
35	菏泽市公共数据开放网	山东省菏泽市	市级	http://hzdata.sd.gov.cn
36	济宁市公共数据开放网	山东省济宁市	市级	http://jindata.sd.gov.cn
37	莱芜市公共数据开放网	山东省莱芜市	市级	http://lwdata.sd.gov.cn
38	聊城市公共数据开放网	山东省聊城市	市级	http://lcdata.sd.gov.cn
39	临沂市公共数据开放网	山东省临沂市	市级	http://lydata.sd.gov.cn
40	日照市公共数据开放网	山东省日照市	市级	http://rzdata.sd.gov.cn
41	泰安市公共数据开放网	山东省泰安市	市级	http://tadata.sd.gov.cn
42	威海市公共数据开放网	山东省威海市	市级	http://whdata.sd.gov.cn
43	潍坊市公共数据开放网	山东省潍坊市	市级	http://wfdata.sd.gov.cn
44	烟台市公共数据开放网	山东省烟台市	市级	http://ytdata.sd.gov.cn
45	枣庄市公共数据开放网	山东省枣庄市	市级	http://zzdata.sd.gov.cn
46	淄博市公共数据开放网	山东省淄博市	市级	http://zbdata.sd.gov.cn

3.3 数据采集及分析

本研究采用网络自动抓取和人工观察相结合的方法采集数据。以 2018 年 4 月 13 日 –2018 年 4 月 18 日为数据采集周期。主体评估分析部分基于截至 2018 年 4 月 18 日从各地平台上所采集的数据，而在对“动态更新”这一指标的评测中则使用了 2018 年 1 月 1 日至 2018 年 4 月 18 日这一时间段内所采集的数据。本研究对采集到的各项指标的数据主要使用描述性统计分析、交叉分析、文本分析等方法进行分析。

4 研究发现

4.1 数据数量

4.1.1 数据集总量 数据集是由数据组成的集合，通常以表格形式出现，每一“列”代表一个特定变量，每一“行”则对应一个样本单位。政府数据开放平台往往以下载或 API 接口的形式开放数据集。个别平台在本研究中未被视作有效的开放数据集，主要有以下 3 类情况：①数据集名称下不存在可直接下载或通过接口获取的数据集；②数据集中仅有 0 –2 行数据的低容量数据集（多为一个数据集分拆出的单行数据，或未整合成一个数据集的单行数据），这类数据的再利用价值很低，不能视作有效数据集；③数据集名称下提供的是网页链接，且链接跳转后出现无法通过下载或接口形式获取的文本内容。开放的有效数据集总量（含直接下载和 API 接口开放）最高的 10 个地方平台如图 1 所示：

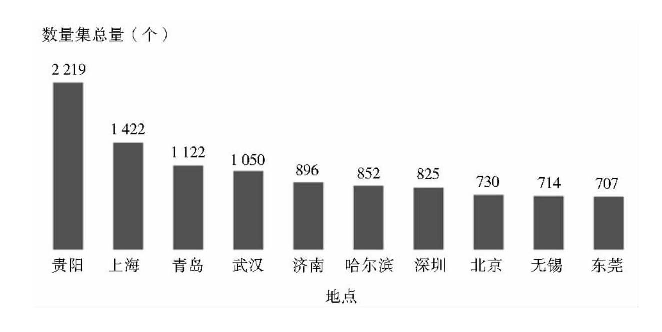


图 1 各地平台上的数据集总量（前 10 名）

4.1.2 数据容量 数据容量是指在各地平台可下载、结构化的数据集中，将字段数（列数）乘以条数（行数）得出的数据总量，用以衡量平台上提供的数据集的实际数据量大小。数据容量排名前 10 的地方平台见图

2。各地平台间数据容量的差距较为明显，排名前 3 的数据容量均超过 8 000 万，但仍有超过三分之一的平台开放的数据容量在 10 万以下。

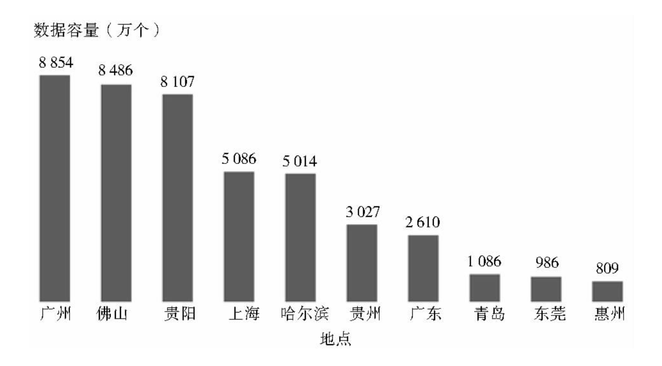


图 2　各地平台上的开放数据容量(前 10 名)

4.2　数据质量

4.2.1　优质数据　优质数据指的是数据量大，社会需求高的数据集。本研究对各地平台上所有可下载的数据集按照数据容量进行排序，共发现了 146 个优质数据集，其分布状况见图 3。在 46 个政府数据开放平台中，17 个平台有优质数据集入选，其他地方平台没有发现优质数据集。表 5 是排名前 10 位的优质数据集名称与其所属平台，这些数据集普遍具有较高的条数、字段数和下载量等，内容上主要和商事主体、药品等相关。

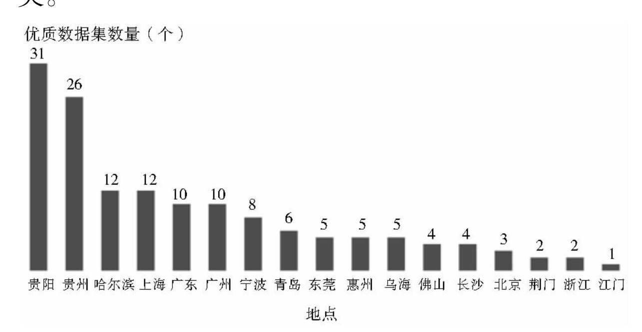


图 3　各地平台的优质数据集数量

表 5　前 10 位优质数据集

序号	数据集名称	所属地方平台	数据容量(个)	条数(条)	字段数(个)	下载量(次)
1	工商登记信息	东莞	46 416 553	2 018 111	23	10 063
2	商事主体个体年报基本信息	广州	34 100 000	1705 000	20	524
3	商事主体基础信息	佛山	8 259 768	458 876	18	365
4	自然人信息	佛山	6 554 376	1 638 594	4	168
5	黑龙江省统一药品信息	哈尔滨	6 106 135	174 461	35	177
6	工程－投标人名单	贵州	4 466 102	235 058	19	－－
7	哈尔滨市个体基本信息	哈尔滨	4 353 930	483 770	9	281
8	贵阳市城镇居民医疗保险药品目录	贵阳	4 294 512	238 584	18	841
9	惠州市工商开业登记信息	惠州	3 359 715	223 981	15	80
10	哈尔滨市商事主体个体年报基本信息	哈尔滨	3 274 398	545 733	6	2 757

4.2.2　无低质数据

（1）无低容量数据。低容量数据是指条数在两行或两行以内的数据集，其原因可能是数据量本身稀少或是数据经统计归总后颗粒度过大，此类数据的再利用价值较低。在 46 个开放数据平台中，近三分之二的平台上存在低容量数据。

（2）无碎片化数据。碎片化数据是指按照时间、行政区划、政府部门等被人为分割的数据集，这些数据集进行整合后将更有利于社会的开发利用。目前大部分的地方开放数据平台均存在碎片化数据。

4.2.3　无问题数据

（1）无重复创建。重复创建是指平台上重复出现标题相同、可下载数据文件相同、且所属主题相同的数据集。在 46 个开放数据平台中，约三分之一的平台存在重复创建问题。

（2）无生硬格式转化。生硬格式转化是指平台将非结构化的 DOC、PDF 等文件中的数据通过生硬方式转化成 XLS、CSV 等机读格式，而数据实质上仍是非结构化的情况。例如，将 WORD 文件中大段的文字贴到 XLS 文件中，将 DOC 格式直接转换成 XLS 格式等。本研究发现有 8 个地方平台存在上述问题。

（3）无无效数据。无效数据是指以下 3 类情况：①数据集名称下没有数据可供获取；②只提供数据链接，无法获取数据集；③数据集下载打开后，里面实际上并不提供数据。在 46 个政府数据开放平台中，超过半数存在无效数据。

4.3　数据标准

4.3.1　开放授权　开放数据应通过数据开放授权协议从法律上保障数据的开放性。目前，各地平台上的数据开放授权通常包含在网站声明、免责条款或服务

协议中。本研究发现，在 46 个政府数据开放平台中，共有 33 个平台配有数据开放授权协议。开放授权协议的内容应明确授予用户免费获取、不受歧视、自由利用、自由传播与分享开放数据的权利。目前仅有 5 个地方平台的授权协议全部明确授予了上述 4 项权利，分别为北京、上海、贵州、广州和贵阳；大部分地区满足了免费获取和不受歧视两项，而其余地区在 4 项指标上未明确提及，或语焉不详。

（1）免费获取。免费获取是指平台在开放授权中明确授予用户免费获取和利用开放数据的权利。政府数据作为公共资源，原则上应免费向社会开放，除非需要对数据进行额外的增值加工和针对少数用户的个性化加工等。目前，各地平台上的相应条款分为“免费且未设时限”和“现阶段免费”两类。其中，只有贵州、贵阳、东莞等地的条款中明确指出数据免费且未设时限，用户可永久无偿获取数据平台所提供的所有数据资源。其他大部分地方平台则在服务协议中提到“现阶段免费”，但设置有模糊的期限或限制，如“保留收费权利” 等表述。

（2）非歧视性。开放授权是指平台明确授予任何用户平等访问、获取、使用和分享开放数据的权利。目前各地平台上的相应条款均明确保障了数据开放的非歧视性，对任何用户都予以平等的数据获取和利用权限，如“用户享有数据资源的非排他使用权”“不受歧视”等表述。

（3）自由利用。开放授权应明确授予用户不受限制地对“开放数据”进行商业和非商业性利用的权利。目前各地平台上相应条款分为“明确允许自由利用”和“未提及可自由利用”两类。其中，只有北京、广州、贵阳、贵州、上海等地明确表示用户可“不受限制地进行商业和非商业性利用”“享有增值利用的权利”或“可自由利用”，其余地方的条款中均未对用户利用数据的权利做出明确说明。

（4）自由传播与分享。开放授权应明确授予用户可自由传播和分享开放数据的权利。目前各地平台上的相应条款分为“可自由传播”“未提及可自由传播”“自由传播受限”3 类。北京、广州、贵阳、贵州、上海等地授予用户享有免费传播现有开放数据的权利。

4.3.2　技术性开放　本研究基于 T. Berners-Lee 提出的开放数据五星标准和其他有关开放数据格式的标准，对各地政府数据开放平台上的数据集的格式标准进行评估。

（1）可机读格式。为方便用户获取和利用数据，数据集应以可机读格式开放，如 XLS、CSV、JSON、XML 等格式。图 4 展示了各地平台上可下载数据集总量与可机读数据集总量对比的前 10 名。总体上，已有 38 个平台开放的数据集基本满足了可机读格式的要求，但也有个别地方平台上出现的数据集为 DOC、PDF、JPG 等不可机读格式。

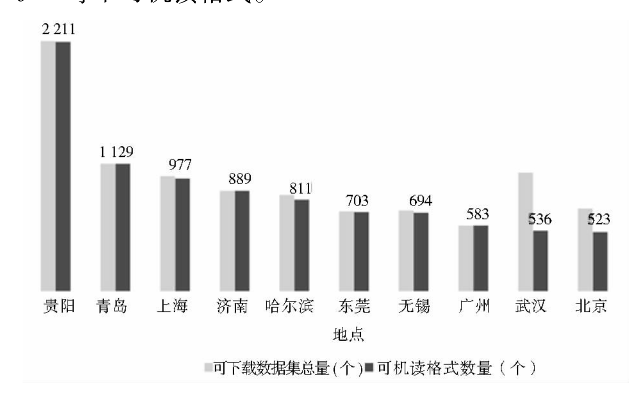


图 4　各地平台可下载数据集与可机读格式数据集数量（前 10 名）

（2）开放格式。开放格式是指可下载数据集应以开放的、非专属的格式提供，任何实体不得在格式上排除他人使用数据的权利，以确保数据无需通过某个特定（特别是收费的）软件或应用程序才能访问。例如 CSV 是开放格式，而 XLS 则不是。图 5 是各地方平台上可下载数据集总量与开放格式总量对比的前 10 名。目前，有 24 个地方平台上提供的数据集全部满足开放格式的标准，其他平台则没有提供任何开放格式的数据集。

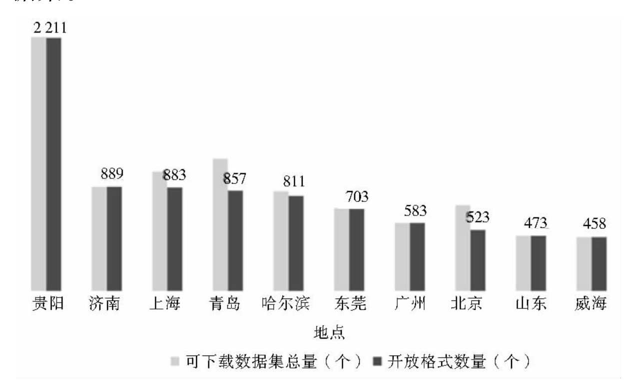


图 5　各地平台可下载数据集与开放格式数据集数量（前 10 名）

（3）RDF 格式。本研究还对 RDF 格式进行了评估，即开放数据五星标准中达到的四星要求。目前，我国仅有贵阳提供了符合 RDF 格式的数据集，共有 216 个 RDF 格式的数据集。

（4）API 接口比例。除了通过直接下载方式提供

数据外,还可通过接口方式使用户实时高效地获取数据,满足其开发应用程序的需求,尤其适合用于开放实时性强、规模大的数据。目前,我国有 16 个地方平台为每个数据集提供了接口,但仍有近三分之一的平台没有提供或仅提供了少量的 API 接口。

4.3.3　元数据完整性　提供元数据有助于数据利用者清楚地了解数据集的内容与背景,从而更好地获取和利用数据。

(1)基本元数据覆盖率。综合梳理我国《政务信息资源编目编制指南(试行)》中关于核心元数据的定义描述、国际开放数据平台上提供的基本元数据条目以及目前我国半数以上的平台已实际提供的元数据条目,本研究确定了以下 12 个条目作为开放数据集基本的元数据条目,包括数据名称、摘要简介、标签关键字、数据主题、数据格式、开放属性、提供单位、发布日期、更新日期、更新频率、数据指标、数据量。

图 6 反映出以上 12 个基本元数据条目在 46 个地方平台的分布情况。目前,46 个地方平台全部都已提供了数据集名称,大多数平台提供了摘要简介、数据提供单位、发布日期、数据主题、数据格式等,而能提供更新频率、数据量、数据指标的地方平台还相对较少。

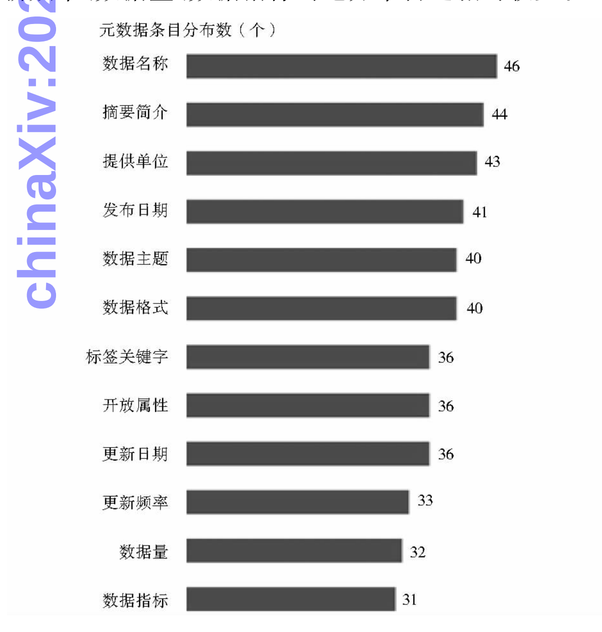


图 6　基本元数据条目在各地平台的分布数

(2)API 描述规范。API 描述有助于数据利用者清楚地了解 API 的具体信息及获取方式,从而更好地调用接口并获取数据。本研究从数据资源描述和数据调用说明两方面评估 API 描述情况。资源描述是指 API 的基本信息,如名称、简介、提供部门、更新时间等;数据调用说明指的是 API 的调用方式、请求地址等信息。在提供 API 接口的 36 个地方平台中,33 个地方平台均提供了资源描述和数据调用说明。

4.4　数据覆盖面

4.4.1　主题覆盖率　提高数据开放的广度和覆盖率有利于数据利用者对来自多种领域的数据进行融合利用。本研究将开放数据主题归纳为经贸工商、交通出行、机构团体、文化休闲、卫生健康、教育科技、社会民生、资源环境、城建住房、公共安全、农业农村、社保就业、财税金融、信用服务共 14 个大类。图 7 体现了各地平台在 14 个主题下所开放的数据集数量。其中,社会民生、经贸工商、教育科技等主题的数据集开放数量最多。在 14 个主题领域中,不同地方平台的主题覆盖情况差异明显,广州和青岛开放的数据集覆盖了全部的 14 个主题。

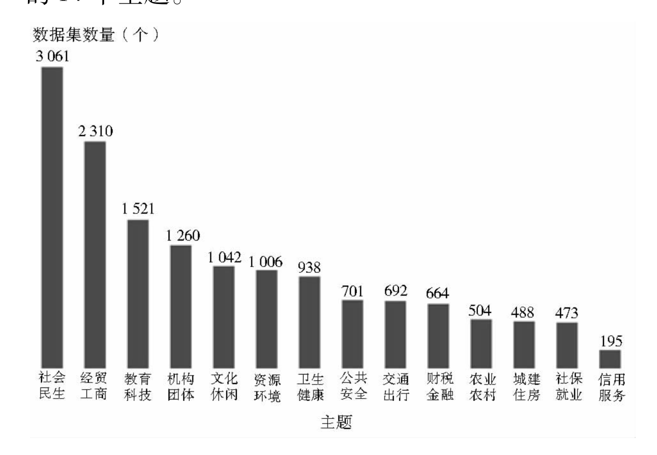


图 7　各主题包含的数据集数量

4.4.2　部门覆盖率　开放数据集的部门覆盖率反映了一个地方政府的各个部门充分参与数据开放工作的程度和数据集来源的全面程度。本研究梳理了各地平台上开放数据集较多的部门作为“主要数据提供部门”。图 8 反映了各地平台上不同类型的政府部门开放的数据集总量,其中统计部门开放的数据最多,其次为教育、交通和卫生部门。然而,统计部门提供的数据多为经过加工归总后的宏观数据,颗粒度较大、数据容量较低,不利于数据被利用和产生价值。进一步分析发现,在统计局数据所占比例较高的地方,其开放数据的容量也普遍较低(见图 9)。

4.4.3　高需求关键词覆盖率　本研究对各地平台上可获取、且下载量最高的前 20% 的数据集名称进行文本分析,发现一批出现频次较高的关键词,反映了各地开放的高需求数据集的内容及其分布。图 10 为其中高频出现的描述性限定词,如“企业”“许可”“建设”和“生产”等。

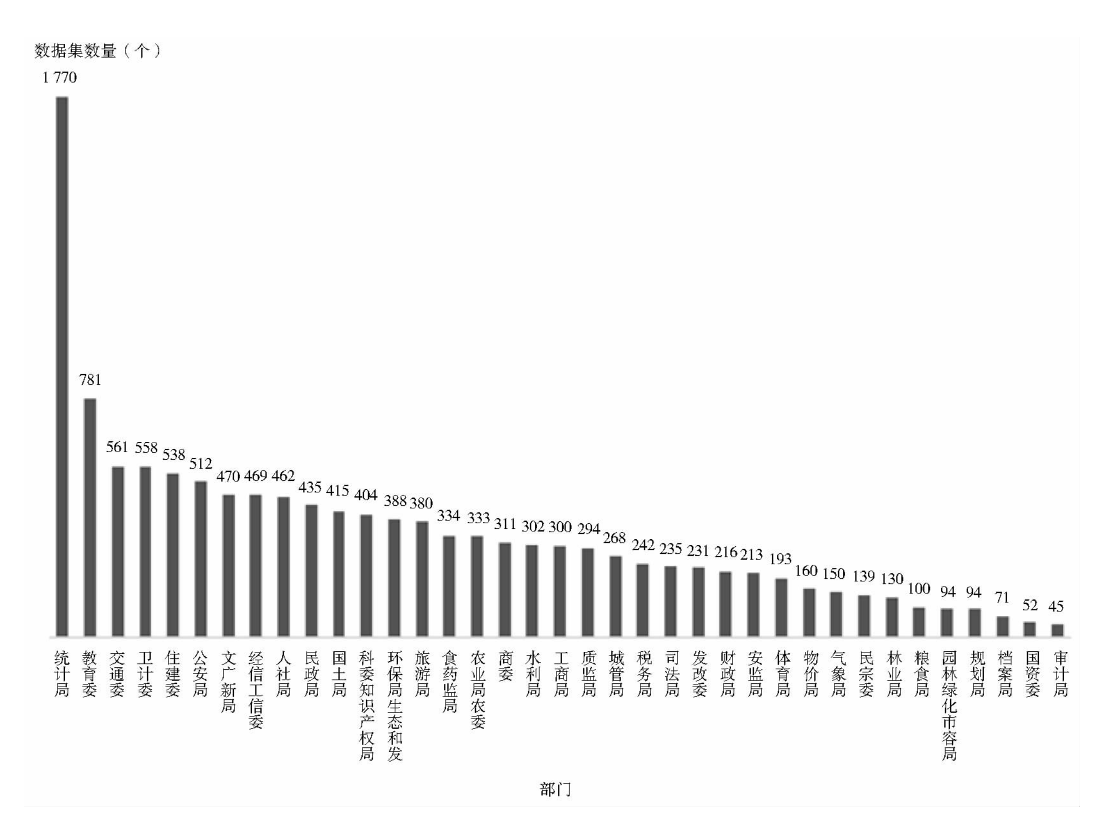


图 8　各地平台主要数据提供部门提供的数据集数量分布

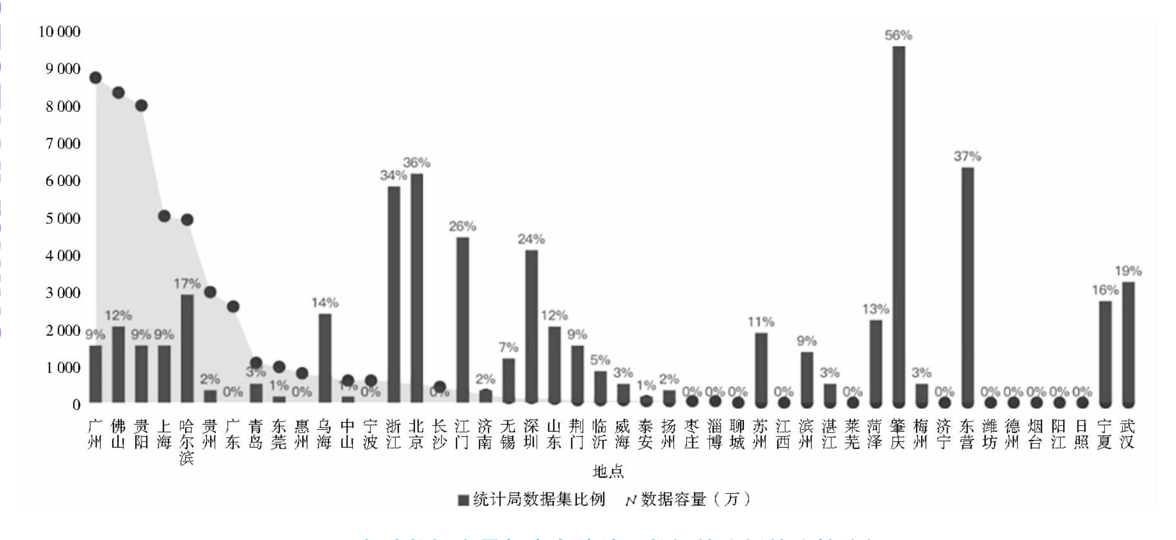


图 9　各地数据容量与来自统计局数据的比例的比较分析



图 10　各地高需求开放数据集名称中出现频次最高的描述性限定词

4.4.4　数据持续性

(1)持续增长。本研究根据各地平台上数据集的创建日期来判断该平台数据集是否持续增长，以季度为时段进行跟踪分析。从平台上线开始，以季度为观测时段的全国各地平台数据集持续增长情况见图 11，颜色区域表示该时段有新增数据集，空白区域则表示该时段无新增数据集。其中，上海市平台从 2012 第四季度发布数据开始，至今数据已保持了 15 个时段的持续增长。2016 年以来各地上线的新平台也大都能保持数据集定期增长。

图 11　各地平台上线时间与数据集持续增长情况

（2）动态更新。该报告跟踪考察了地方数据平台在 2018 年 1 月至 2018 年 4 月期间更新的数据集数量，情况见图 12。贵阳平台在该时段内更新的数据集数量最多，超过 1 000 个，但有少数地方平台完全没有更新数据。

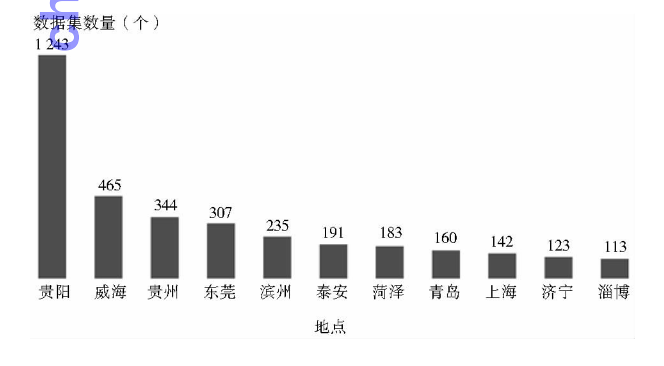


图 12　各地平台数据集动态更新数量（前 10 名）

（3）历史存档。历史存档是指平台将历史上不同时间更新的多个批次的数据同时留存在平台上供用户下载，有利于数据利用者按时间线索来获取和利用历史数据。图 13 反映了数据历史存档的平台分布情况。目前有上海、广东、广州等 15 个地方平台实现了数据历史存档。

5　研究结论

5.1　从国家层面深入到地方层面

目前，针对政府数据开放已发布了多个权威性的国际评估报告，其中影响力最大的两个是万维网基金会发布的“开放数据晴雨表”和英国开放知识基金会组织发布的“全球开放数据指数”。“开放数据晴雨表”是由万维网基金会开展的全球性评估项目，于 2013 年启动，采用专家调查、辅助数据、同行评估、定量数据和定性评估结合的方式，每两年左右从“准备度”“执行”和“产生的影响”3 个层面对各国政府数据开放 进行评估。2016 年，“开放数据晴雨表”对 115 个国家进行了评估，中国排名第 71 位。“全球开放数据指数”是由开放知识基金会在全球范围内进行的评估项目，主要对各国开放的关键数据集进行评估，采用滚雪球抽样、志愿者问卷调查和访谈的持续性众包、专家评估、同行评估等方式进行。2015 年该指数评估了 122 个国家和地区，中国排名第 93 名。

从国际政府数据开放评估报告的结果可见：一方面，我国政府数据开放在国际上仍处于靠后的位置，有

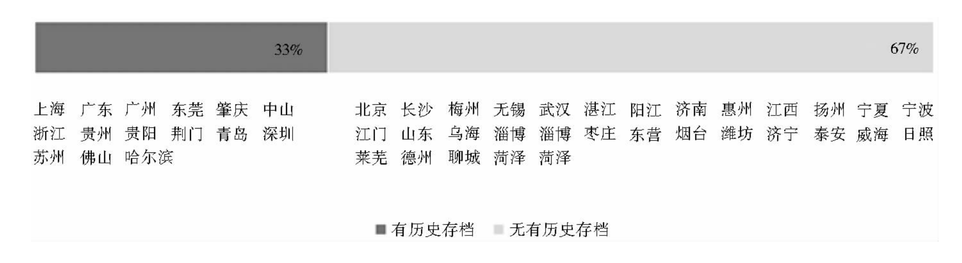


图 13　实现数据历史存档的平台分布

待提升。另一方面，国际数据开放评估报告的评估体系和方法也并不完全适用于中国政府数据开放的实际现状和发展阶段。例如，针对法律政策和体制机制等方面的评估指标还无法直接运用于中国情景；针对数据开放后产生的利用效果和影响的评估对于中国目前的数据开放现状而言还有些超前，而且这些评估报告也未能对中国地方层面政府数据开放的现状进行评估。

因此，针对过去国际评估未能触及到的地方政府层面的数据开放，本研究构建了一个更为符合中国实际的评估框架，并开展了实际测评。该框架聚焦于目前我国地方政府数据开放的核心——数据层面，并结合我们地方政府数据开放的实际发展阶段和社会需求，重点针对数据质量、数据标准、数据可持续性、数据数量和数据覆盖面等维度进行评估。与国际上现有的评估体系相比，这一框架更有助于真正推动中国地方政府数据开放工作的发展。同时，相比于之前国内学者已开展的地方评估，本研究的评估样本更为全面，覆盖了我国 46 个省级和地级政府，并且在评估指标上也更加聚焦。

5.2　数据数量稳步增长，但数据集容量偏低

研究发现，目前我国各地开放数据集总量稳步增长，其中贵阳、上海、青岛、武汉等地的数据集总数已突破 1 000 个，但仍有 20 多个地方的开放数据集总数还不足 200 个。另外需要引起重视的是，各地开放的数据集数据容量整体偏低，多数地方开放的数据集行数与列数过少，这类数据集无法被有效利用。

5.3　数据质量参差不齐，问题数据普遍存在

部分地区已上线了一批高容量、高需求的优质数据集，但大多数地方的数据集质量参差不齐，各地普遍存在低容量碎片化的低质数据或重复创建、格式生硬转化和无效的问题数据。这些低质量数据和问题数据很难被再次利用并产生相应价值，使数据开放流于形式。

5.4　数据标准有待规范和提升

在数据的法律性开放上，少部分地方平台的授权协议中明确授予了用户免费获取、不受歧视、自由利用、自由传播和分享数据的权利，但许多地方的政府数据开放平台仍未提供明确充分的数据开放授权。

在数据的技术性开放上，各地平台上可机读、非专属、以接口形式提供的数据集比例稳步增长。贵阳市在全国率先提供了 RDF 格式的数据集。然而，全国各地仍然存在很多不符合开放数据格式标准的数据集。

在元数据完整性上，目前大多数地方平台都能提供基本的元数据，但各地情况参差不齐，普遍缺少数据集的发布时间、更新时间、数据指标和数据量等条目。

5.5　数据覆盖面较低，以统计数据为主

在主题覆盖率上，目前各地平台上提供最多的数据集主题是社会民生和经贸工商，但不同地方平台的主题覆盖情况不一。

在部门覆盖率上，各地平台上开放数据集最多的部门是统计局。统计局发布的多为二手的、经过加工归总的数据集，其再利用价值低于来自业务部门的、一手的、原始的数据集。全国仍有接近半数的地方政府各部门参与开放数据的程度不到一半。

在关键词覆盖率上，各地开放的高需求数据集名称中出现“企业”“许可”“建设”和“生产”的频次最高，但有少数城市的覆盖率不到三成。

5.6　数据可持续性偏低，日常更新和增长不足

少数地方能基本保持新增数据集持续增长与存量数据集动态更新，但不少地方平台出现数据集增长间歇或停滞，真正实现存量数据动态更新的比例仍然偏低。此外，仅有不到一半的地方将历史上多个批次的数据留存在平台上供用户获取。多数地方平台对开放数据持续运维与持续更新的重视程度不够。

6　对策建议

6.1　提升数据数量

既要提升开放数据集的总体数量，更要注重提升数据集的数据容量，也就是要提升开放数据集中的字

段数(列数)和条数(行数)。

6.2 提高数据质量

开放价值密度高,社会需求高的优质数据集,随意发布一些易于发布的、低密度、碎片化、有问题的数据并不会创造价值。开放数据应多从用户的实际需求而非政府部门的自我判断出发,定期向数据利用者征集需求和建议,有针对性地开放社会真正有需求的、能解决问题和创造价值的优质数据集,并确保数据的完整性、准确性和适用性。

6.3 规范数据标准

6.3.1 法律性开放　为开放数据提供授权协议,明确授予用户免费获取、不受歧视、自由利用、自由传播与分享开放数据的权利,并进一步探索分级分类的方式,对不同的数据集配备不同内容的授权协议。

6.3.2 技术性开放　基于开放数据的基本原则和标准,开放完整的、原始的、可机读的、开放格式的、结构化的、电子化的数据集,让用户能把数据真正用起来。依据 T. Berners-Lee 提出的开放数据五星标准,现阶段我国大部分地区的开放政府数据已符合三星标准,下一步各地政府数据开放应向四星标准 RDF 格式迈进,并继续向五星标准发展,使数据之间实现关联。

除了确保数据可被直接下载,政府数据开放平台还应对数据规模大、动态实时性强、处理要求高的数据通过 API 接口方式进行开放。还要为 API 接口提供规范描述,包括资源描述和数据调用说明,帮助数据利用者了解 API 的具体信息及获取方式,从而更好地调用接口并获取数据。

6.3.3 元数据完整性　平台在开放数据集的同时还应提供全面的元数据信息,以帮助数据利用者清楚地了解数据集的内容与背景,从而更好地理解和利用数据。元数据条目可包括数据名称、摘要简介、标签/关键字、数据主题、数据格式、开放属性、提供单位、发布日期、更新日期、更新频率、数据量和字段名称等。

6.4 扩大数据覆盖面

尽可能覆盖重点开放领域的关键数据集,提升数据的广度、丰富度和针对性,使数据利用者可充分获取和整合多种来源的数据,进行深度挖掘和利用。政府需着力提高各个业务部门参与开放数据的程度,而不是将开放数据的重点部门都放在统计部门上。

6.5 保持数据可持续性

开放政府数据是一项持续性和常态化的工作,数据集在开放后还需持续更新和增加。只有源源不断的数据供给,才能激发数据利用的活力,满足社会对开放数据日益增长的需求。政府应建立长效工作机制,确保开放数据集存量动态更新,增量持续不断,并将不同时间开放的历史数据留存在平台上供数据利用者继续下载利用。

参考文献:

[1] 郑磊,开放政府数据的价值创造机理:生态系统的视角[J]. 电子政务,2015(7):2－7.

[2] The annotated 8 principles of open government data. Open government data principles[EB/OL]. [2018－05－04]. https://public.resource.org/8_principles.html.

[3] Open definition. The open definition[EB/OL]. [2018－05－04]. https://opendefinition.org.

[4] The World Bank. Open data essentials[EB/OL]. [2018－05－04]. http://opendatatoolkit.worldbank.org/en/essentials.html.

[5] Linked Data[EB/OL]. [2017－08－15]. https://www.w3.org/DesignIssues/LinkedData.html.

[6] Open Data Chapter. Principles[EB/OL]. [2018－05－04]. https://opendatacharter.net/principles.

[7] OVIEDO E, MAZON J N, ZUBCOFF J J. Towards a data quality model for open data portals[C]//Computing Conference (CLEI), 2013 XXXIX Latin American computing conference. Naiguata: IEEE, 2013: 1－8.

[8] VISCUSI G, SPAHIU B, MAURINO A, et al. Compliance with open government data policies: an empirical assessment of Italian Local Public Administrations[J]. Information polity, 2014, 19(3/4):263－275.

[9] LOURENÇO R P. An analysis of open government portals: a perspective of transparency for accountability[J]. Government information quarterly, 2015, 32(3): 323－332.

[10] BELLO O, AKINWANDE V, JOLAYEMI O, et al. Open data portals in Africa: an analysis of open government data initiatives[J]. African journal of library, archives & information science, 2016, 26(2): 97－106.

[11] The Gov Lab. Open data definitions[EB/OL]. [2017－08－15]. http://odimpact.org/resources.html.

[12] SUSHA I, ZUIDERWIJK A, JANSSEN M, et al. Benchmarks for evaluating the progress of open data adoption: usage, limitations, and lessons learned[J]. Social science computer review, 2015, 33(5): 613－630.

[13] 夏义堃. 国际组织开放政府数据评估方法的比较与分析[J]. 图书情报工作,2015,59(19):75－83.

[14] 郑磊,关文雯. 开放政府数据评估体系、指标与方法研究[J]. 图

书情报工作,2016,60(18):43 - 55.
[15] 郑跃平,刘美岑. 开放数据评估的现状及存在问题——基于国外开放数据评估的对比和分析[J]. 电子政务,2016(8):84 - 93.
[16] 陈美. 开放政府数据价值评估: 进展与启示[J]. 情报杂志,2017(11):92 - 98.
[17] 韦忻伶,安小米,李雪梅,等,开放政府数据评估体系述评: 特点分析[J]. 图书情报工作,2017,61(18):119 - 127.
[18] 郑磊,高丰. 中国开放政府数据平台研究:框架、现状与建议[J]. 电子政务,2015(7):8 - 16.
[19] 郑磊,熊久阳. 中国地方政府开放数据研究:技术与法律特性[J]. 公共行政评论,2017(1): 53 - 74.
[20] 夏义堃. 国际比较视野下我国开放政府数据的现状、问题与对策[J]. 图书情报工作,2016,60(7):34 - 40.
[21] 赵继娣,张罕仑. 地方政府数据开放成效评价研究——以上海市为例[J]. 电子政务,2017(9): 11 - 21.
[22] 沈晶,韩磊,胡广伟. 政府数据开放发展速度指数研究——基于我国省级政府数据开放平台的评估[J]. 情报杂志,2018(8):1 - 8.
[23] 海伦,邓崧. 基于数据包络法的城市政府数据开放平台效率评估[J]. 电子政务,2018(8):112 - 118.

作者贡献说明:
郑磊:负责梳理文献,确定评估框架和方法,撰写分析文字;
吕文增:负责抓取和分析数据,制作图表。

Assessing Open Data at Local Government Level: Framework and Findings

Zheng Lei Lü Wenzeng
School of International Relations and Public Affairs, Fudan University, Shanghai 200433

Abstract: [**Purpose/significance**] This paper attempts to construct an assessment framework on open government data at local government level, evaluate the data dimension on existing local government open data platforms in China and put forward suggestions to foster the opening of local government data. [**Method/process**] Based on the definitions, principles and standards of open data, learning from international open data assessment frameworks, taking in consideration of the policy requirement and development status of open data practices in China, this paper constructs a systematic, multi-dimensional and operable assessment framework, and carries out an actual assessment on forty-six local government open data platforms in China. [**Result/conclusion**] The study finds out a number of problems with regard to the quantity, quality, standard, coverage and sustainability of open data on local government platforms in China.

Keywords: China local government open data assessment framework